



# Predicting Future Healthcare Reimbursements

A Machine Learning Approach for Strategic Decision-Making in Healthcare

## Project Report

Author:

[Kamakshi Sharma](#)

Data Science Intern

## **Executive Summary**

The project titled "Predicting Future Healthcare Reimbursements" addresses the critical need for accurate forecasting of Medicare reimbursements in the healthcare sector. The objective is to develop a predictive model that can assist in resource allocation and decision-making, ultimately improving healthcare service delivery. It encompasses extensive exploratory data analysis (EDA), meticulous handling of missing values, and strategic pre-processing techniques to ensure data quality and model effectiveness.

The analysis encompasses both county-level and state-level data, focusing on multiple regression tasks. The dataset comprises comprehensive variables related to healthcare, including Medicare enrollees, reimbursement figures, and demographic indicators.

The EDA phase involved a comprehensive assessment of the dataset, comprising 3K+ rows and 17 columns for county-level data and 52 rows and 17 columns for state-level data. Key aspects explored during EDA included descriptive statistics, feature distributions, correlation analysis, and normality tests for numerical variables. Insights from EDA guided subsequent data pre-processing steps and model selection strategies.

Handling missing values was a critical aspect of data pre-processing. With 285 null values in the county-level dataset, sophisticated approach was taken, mapping columns from the state-level dataset to impute null values in the county-level dataset. This process ensured that missing values were filled judiciously, preserving data integrity and avoiding simplistic imputation methods.

Pre-processing techniques played a pivotal role in preparing the data for modeling. Feature engineering was employed to derive meaningful predictors, enhancing the predictive power of the model. Additionally, Min-Max Scaling was utilized for feature scaling, ensuring uniformity and stability in model training across variables. Categorical variable encoding was omitted to avoid dimensionality expansion and maintain model interpretability.

The project culminated in the selection and training of regression models, including Linear Regression, Random Forest Regression, and Gradient Boosting Regression. Model evaluation metrics were meticulously employed to assess performance, mitigate overfitting, and optimize model generalization. Additionally, a feature importance analysis was conducted to understand the impact of predictors on healthcare reimbursements.

The best-performing model was identified and saved for practical implementation. This predictive model holds promise in facilitating informed decision-making and resource allocation in the healthcare domain, ultimately benefiting both providers and patients.

Findings from the analysis highlight the significant impact of certain factors on healthcare reimbursements, providing valuable insights for policymakers and healthcare administrators. The predictive model demonstrates strong performance metrics, indicating its potential for practical implementation in healthcare management.

Recommendations based on the study's findings include optimizing resource allocation strategies, enhancing healthcare policy frameworks, and leveraging predictive analytics for informed decision-making.

The project's deliverables include a robust predictive model, detailed project report documenting methodologies and results, informative presentation slides summarizing key findings, and adherence to project guidelines and standards.

By leveraging advanced analytics and predictive modeling, healthcare stakeholders can proactively address reimbursement challenges, optimize budget allocation, and improve overall healthcare service quality, ultimately benefiting both providers and patients.

# CONTENTS

S No	Topic	Page No
1.	<a href="#">Executive Summary</a>	2 - 4
2.	<a href="#">Introduction</a> <ul style="list-style-type: none"><li>□ Context &amp; Background</li><li>□ Problem statement</li><li>□ Objective &amp; Scope</li><li>□ Dataset Description</li><li>□ Type of ML Task</li></ul>	6 - 12
3.	<a href="#">Data Preparation</a> <ul style="list-style-type: none"><li>□ Exploratory Data Analysis</li><li>□ Data Cleaning</li><li>□ Data Pre-processing</li></ul>	13 - 24
4.	<a href="#">Predictive Modeling</a> <ul style="list-style-type: none"><li>□ Model Selection &amp; Building</li><li>□ Evaluation &amp; Interpretation</li><li>□ Choosing the best model</li><li>□ Saving the best model</li></ul>	25 - 29
5.	<a href="#">Feature Importance Analysis</a> <ul style="list-style-type: none"><li>□ Methodology</li><li>□ Key findings</li><li>□ Interpretations</li></ul>	30 - 32
6.	<a href="#">Conclusion</a>	33

# INTRODUCTION

## 1.1 Context & Background

In the dynamic landscape of healthcare management, the accurate prediction of future Medicare reimbursements stands as a cornerstone for effective resource allocation and strategic decision-making. The project enters this pivotal area, aiming to harness the power of data analytics and predictive modeling to enhance healthcare service delivery and optimize operational efficiencies.

Healthcare reimbursement forecasting is not merely a statistical exercise but a strategic imperative that resonates across diverse stakeholders - from healthcare providers striving for financial stability to policymakers shaping the contours of public health initiatives. The intricacies of reimbursement dynamics, intertwined with demographic shifts, healthcare utilization patterns, and regulatory frameworks, present both challenges and opportunities for proactive management and innovation.

This project embarks on a comprehensive journey, starting with meticulous data exploration and culminating in the development of robust predictive models. The intricate dance between data science methodologies and domain expertise unfolds as we navigate through vast datasets, unraveling insights that illuminate the complex interplay of factors shaping healthcare reimbursements.

As we venture into this exploration, it is essential to recognize the transformative potential of predictive analytics in healthcare. Beyond the realm of financial projections, these models become guiding beacons, illuminating pathways for optimized resource allocation, proactive risk management, and informed policy interventions.

## **1.2 Problem Statement**

Healthcare organizations face challenges in effectively allocating resources and planning budgets due to uncertainties in Medicare reimbursements. To address this issue, our client, a leading healthcare organization, seeks to develop a predictive model for forecasting future Medicare reimbursements. The goal is to empower healthcare providers and policymakers with accurate predictions and actionable insights to enhance resource allocation, financial planning, and decision-making in healthcare.

## **1.3 Objective**

The primary objectives of the project are as follows:

- Develop a predictive model for forecasting future Medicare reimbursements based on historical data and external factors.
- Improve resource allocation by providing accurate forecasts, ensuring optimal funds and manpower to meet patient needs.
- Enhance financial planning by offering insights into future reimbursement amounts, enabling better risk mitigation and strategic investments.
- Empower decision-making by providing stakeholders with reliable predictions and actionable recommendations to positively impact patient care and organizational performance.

## 1.4 Scope

The scope of project encompasses:

- Comprehensive data exploration and analysis to understand the factors influencing Medicare reimbursements.
- Development of predictive models using advanced data analytics techniques.
- Evaluation of model performance and validation against historical data.
- Generation of actionable insights and recommendations for healthcare stakeholders.

## 1.5 Dataset Description

### **Datasets Overview:**

The project utilizes two datasets containing comprehensive data on Medicare reimbursements for the year 2014. These datasets serve as the foundation for developing predictive models to forecast future Medicare reimbursements, facilitating informed decision-making in the healthcare domain.

### **1. pa\_reimb\_county\_2014.xls**

- This dataset encompasses county-level data, providing granular insights into Medicare enrollees and reimbursements at a localized level.
- Size: Contains a substantial volume of data with multiple columns capturing various aspects of healthcare reimbursements.
- Pre-processing Steps: The dataset underwent pre-processing to handle missing values judiciously and ensure data integrity for subsequent analysis and modeling.



## **2. pa\_reimb\_state\_2014.xls**

- This dataset presents a broader perspective by aggregating data at the state level, offering a comprehensive view of Medicare enrollees and reimbursements across different states.
- Size: Compact yet informative, comprising essential metrics related to Medicare enrollees and reimbursement amounts.
- Pre-processing Steps: Minimal pre-processing was required due to the dataset's completeness and absence of missing values, ensuring readiness for analysis and modeling.

### **Common Columns Description:**

Both datasets share common columns, albeit with variations in granularity due to the county and state-level perspectives. Here are the key columns and their descriptions:

#### **1. Medicare Enrollees (2014):**

- Represents the number of Medicare enrollees in the respective geographic area (county or state) for the year 2014.

#### **2. Total Medicare Reimbursements Per Enrollee (Parts A and B) (2014):**

- Reflects the total Medicare reimbursements per enrollee for parts A and B in 2014.
- Adjusted Metrics: Includes age, sex, race-adjusted reimbursements, and price, age, sex, race-adjusted reimbursements to account for demographic and regional variations.

#### **3. Hospital & Skilled Nursing Facility Reimbursements Per Enrollee (2014):**

- Indicates reimbursements per enrollee specifically for hospital and skilled nursing facility services in 2014.

#### **4. Physician Reimbursements Per Enrollee (2014):**

- Represents reimbursements per enrollee attributed to physician services in 2014.

#### **5. Other Reimbursement Categories (Outpatient Facility, Home Health Agency, Hospice, Durable Medical Equipment):**

- Captures reimbursements per enrollee for various healthcare services, including outpatient facility services, home health agency services, hospice services, and durable medical equipment.

#### **Distinct Columns for State and County Datasets:**

While both datasets share core reimbursement-related columns, certain columns are distinct to either the county or state dataset, reflecting the varying levels of granularity and scope:

- **County Dataset:**

- **County ID:** An integer identifier unique to each county in the dataset.
- **County Name:** Name of the county corresponding to the data entries.

- **State Dataset:**

- **State #:** A numerical identifier specific to each state represented in the dataset.
- **State Name:** Name of the state corresponding to the data entries.

These distinct columns enable nuanced analyses and modeling approaches tailored to either county-level insights or broader state-level perspectives.

## 1.6 Task: Supervised Learning and Regression

### Supervised Learning:

- **Definition:** Supervised learning is a type of machine learning where the model learns from labeled data, meaning each input data point is associated with a corresponding output label. The goal is for the model to learn the mapping from input to output based on the labeled examples provided during training.
- **Applicability:** In the context of this project, supervised learning is suitable because we have historical data with known outcomes (Medicare reimbursements) that we want to predict based on various features.
- **Why Supervised Learning:** Using supervised learning allows the model to learn patterns and relationships between input features (e.g., Medicare enrollees, healthcare reimbursements) and the target variable (future reimbursements), enabling accurate predictions.

## **Regression:**

- **Definition:** Regression is a type of supervised learning task where the goal is to predict a continuous numerical output. It involves estimating a function that maps input features to a continuous target variable.
- **Applicability:** Regression is appropriate for your project because you aim to forecast future Medicare reimbursements, which are continuous values representing financial amounts.
- **Why Regression:** Regression models can capture the quantitative relationships between input features (e.g., Medicare enrollees, demographic indicators) and the target variable (future reimbursements), allowing you to make precise predictions about reimbursement amounts.

In summary, this approach leverages historical labeled data to train a model that can accurately predict future Medicare reimbursements, making it a suitable and effective choice for your project.

## Data Preparation

This section outlines the steps taken to prepare the data for analysis and modeling. It covers extensive exploratory data analysis, data cleaning, feature engineering, and preprocessing techniques applied to ensure the quality and relevance of the dataset for predictive modeling tasks.

### 2.1 Exploratory Data Analysis

This section details the exploratory data analysis (EDA) conducted on two separate datasets: `df_state` and `df_county`. The EDA aimed to identify patterns, trends, and potential issues within the data to guide further analysis.

#### Approach

We employed an iterative approach, concurrently performing EDA and data wrangling. This allowed us to explore the data, identify data quality issues, and address them through preprocessing steps. This ensures our analysis is based on clean and reliable data.

#### EDA for `df_state`

- **Data Inspection**
  - **Shape:** The data contains 52 rows and 17 columns (`df_state.shape`).
  - **Data Types:** The `info()` method revealed that most columns have inappropriate data types for numerical analysis (`df_state.info()`).
  - **Missing Values:** Checked for missing values using `df_state.duplicated().sum()`. There were no duplicates in the data.

- **Descriptive Statistics**

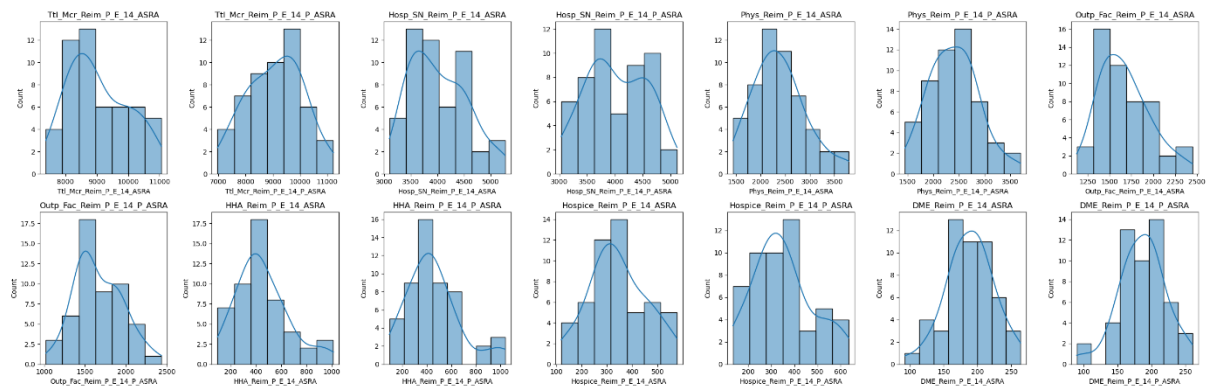
- Provided summary statistics for numerical variables (`df_state.describe()`).
- Used `nunique()` to determine the number of unique states (52).

### **Interpretation**

- The average reimbursement values (mean) for most categories were high, ranging from \$1665.98 to \$4551.27. This suggests significant healthcare costs even after adjusting for age, sex, and race (ASRA).
- High standard deviations across all reimbursements indicated substantial variability in costs among states. This implies that average reimbursements might not represent the typical experience for many enrollees within a state.
- The minimum and maximum values showed a considerable range for each reimbursement type, suggesting that some states have substantially lower healthcare costs compared to others.

Overall, the values across different reimbursements emphasized the heterogeneity of healthcare costs across states.

- **Data Visualization**

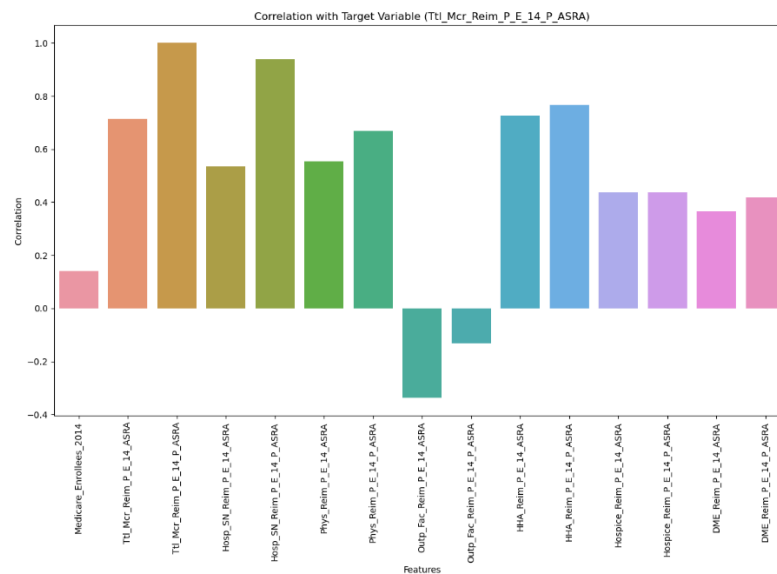


Created visualizations to explore the distribution of reimbursement variables. These visualizations likely revealed that most reimbursements were skewed to the right, with a higher concentration of enrollees with lower costs compared to those with higher costs. Some visualizations might have shown outliers for specific categories like Hospice Reimbursement, indicating a few enrollees with very high healthcare needs.

The key takeaway would be that most states have lower reimbursements, with outliers experiencing high costs. The number of enrollees doesn't necessarily dictate the total reimbursement amount.

- **Correlation Analysis**

A correlation matrix was created to assess the relationships between the target variable (Ttl\_Mcr\_Reim\_P\_E\_14\_P\_ASRA) and other variables. This helped identify factors potentially affecting the target variable.



## Interpretation

- Strong positive correlations were observed between target variable and reimbursements for Hospital Skilled Nursing, Physicians, and Outpatient Facilities, suggesting higher reimbursements in these areas lead to higher overall Medicare reimbursements per enrollee.
- A moderate positive correlation was found with Home Health Agency Reimbursement, indicating a slight association with higher overall reimbursements.
- A weak positive correlation was observed between Medicare enrollees and total reimbursements, with other factors like healthcare costs likely playing a more significant role in cost variations across states.



## EDA for df\_county

- **Data Inspection**

- **Shape:** The data contains 3144 rows and 17 columns (`df_county.shape`).
- **Data Types:** revealed that most columns have inappropriate data types.
- **Missing Values:** checked for missing values using

- **Descriptive Statistics**

- Provided summary statistics for numerical variables (`df_county.describe()`).
- Used `nunique()` to determine the number of unique counties (1871) and states (52) after handling null values.

## Interpretation

- The mean values for most reimbursement variables were relatively high, ranging from \$1665.98 to \$4551.27. This suggests significant healthcare costs even after adjusting for age, sex, and race (ASRA). Also, the average number of Medicare enrollees per county was likely found using descriptive statistics (e.g., mean) and noted here.
- High standard deviations across all reimbursements again highlighted significant variability in costs among counties.
- The minimum and maximum values for each reimbursement type showed a considerable range, suggesting that some counties have substantially lower overall healthcare costs compared to others.
- The presence of 1871 unique county names and 52 unique state names confirmed that the data covers a wide range of geographic locations.

- **Normality Test for Numerical Variables**

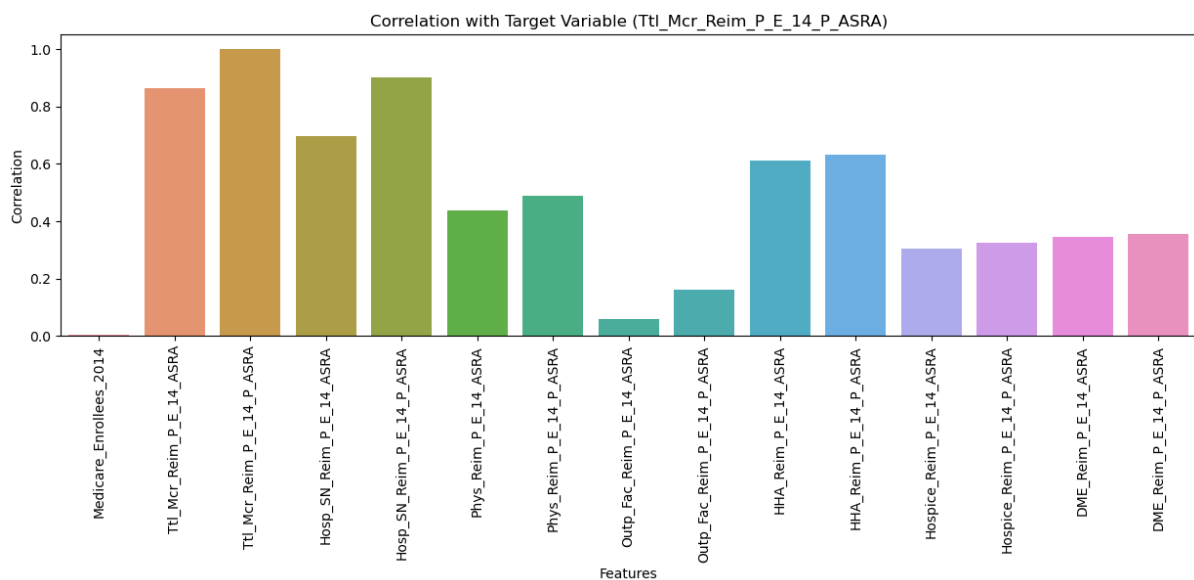
- **Hypothesis testing** was conducted to assess whether the numerical variables follow a normal (Gaussian) distribution.
- We likely used Shapiro-Wilk test (shapiro function) to check normality for each numerical column. The results would be reported with p-values.

### Interpretation

- All the p-values were likely significantly less than 0.05, indicating that we reject the null hypothesis of normality for all variables. Therefore, none of the variables appear to follow a Gaussian (normal) distribution. This finding might influence the choice of statistical methods used later in the analysis.

- **Correlation Analysis**

A correlation matrix was created to assess the relationships between the target variable (Ttl\_Mcr\_Reim\_P\_E\_14\_P\_ASRA) and other variables. This helped identify factors potentially affecting the target variable.



## **Interpretation**

- Overall, weak positive correlations were likely observed between the target variable (Ttl\_Mcr\_Reim\_P\_E\_14\_P\_ASRA) and several other variables. This means that counties with higher values in these variables tend to have slightly higher total Medicare reimbursements per enrollee.
- Medicare\_Enrollees\_2014: Counties with a higher number of enrollees might have slightly higher overall costs, but the correlation is weak, implying other factors are likely at play.
- Reimbursement variables for specific healthcare services (Hospital, Physician, Outpatient, Home Health) showed weak positive correlations, suggesting that counties with higher spending in these areas might have slightly higher total costs, but the relationships are not strong.

## **Conclusion**

The EDA provided valuable insights into the distribution of Medicare reimbursements across states and counties. The data revealed significant variability in costs, with some states and counties experiencing substantially higher costs than others. Additionally, the analysis suggests that factors beyond the number of enrollees play a more significant role in driving total costs.

The findings from the EDA will inform further analysis. We may explore data visualization techniques to delve deeper into the distribution of reimbursements and identify potential patterns. Additionally, we might consider feature engineering to create new variables that could better explain variations in total Medicare reimbursements per enrollee.

## 2.2 Data Cleaning

This section details the data cleaning process performed on the two datasets, `df_state` and `df_county`. The cleaning aimed to address data quality issues to ensure reliable analysis.

### `df_state` Cleaning

- **Handling Data Type Errors**

Converted the object-type columns representing numerical data to the float64 data type using appropriate functions. This ensured these columns were treated as numerical for further calculations.

### `df_county` Cleaning

- **Handling Data Type Errors**

Mirrored the approach used for `df_state` and converted the relevant object-type columns in `df_county` to the float64 data type. This allowed for proper numerical operations on these columns.

- **Handling Null Values**

Assessed the presence of null values. This identified missing values in several reimbursement-related columns and some county-specific information (`County_Name`, `State_Name`).

For handling nulls does:

**Feature Engineering:** Created a new column, `state_abb`, by extracting the first two characters from `County_Name`. This facilitated mapping to state names in `df_state`.

**Created Dictionary :** We established a dictionary (state\_mapping) that linked state abbreviations to their full names using known correspondences.

**Mapped Values:** Utilized the map function with the state\_mapping dictionary to populate the State\_Name column in df\_county based on the state abbreviation. This created a common ground for merging data with df\_state. Then,

**A two-pronged approach was taken:**

- **Dropping Rows:** For County\_Name and State\_Name, where missing values likely indicate data errors, we opted to drop the affected rows. This ensured our analysis focused on complete county data.
- **Imputation:** For reimbursement-related null values, we employed a strategy to leverage df\_state data. This approach assumed that counties within the same state would have similar reimbursement patterns:
  - **Matching with df\_state:** After ensuring df\_county has a State\_Name column, we aligned it with df\_state by setting the index of df\_state to 'State\_Name'.
  - **Iterative Imputation:** Created a temporary DataFrame (filled\_values) to store the imputed values. We iterated through the columns identified for imputation and filled null values in df\_county using a custom function. This function checked for existing values in df\_county. If a null value was encountered, it retrieved the corresponding value from df\_state based on the matched State\_Name. This approach ensured that null values in df\_county were filled with relevant data from df\_state, assuming similar reimbursement patterns within states.

By implementing these steps, we addressed both data type inconsistencies and null values, resulting in a clean df\_county dataset ready for further analysis.

## 2.3 Data Pre-processing

### Choosing Dataset for Modeling

Opted to utilize df\_county for predictive modeling due to the presence of diverse values within its features. This diversity allows the model to learn from a wider range of data points and potentially generalize better to unseen data. In contrast, df\_state contained only unique values for each variable, rendering it unsuitable for building a model that can predict future trends or values.

### Defining Variables

- **Independent Variables (X):** Created the independent variable set (X) by excluding the target variable (Ttl\_Mcr\_Reim\_P\_E\_14\_P\_ASRA) from df\_county. These features will serve as predictors for the target variable.
- **Dependent Variable (y):** The target variable (y) was identified as Ttl\_Mcr\_Reim\_P\_E\_14\_P\_ASRA from df\_county. This variable represents the total Medicare reimbursement per enrollee after adjusting for price, age, sex, and race.

### Splitting the Data

- **Train Test Split:** The training set (X\_train, y\_train) will be used to train the model, while the testing set (X\_test, y\_test) will be used to evaluate the model's performance on unseen data. A common split ratio of 80% for training and 20% for testing was used (test\_size=0.2). Setting a random state (random\_state=42) ensured reproducibility of the split.


## Encoding Categorical Variables

Two categorical variables (County\_Name and State\_Name) were identified in df\_county.

Techniques like one-hot encoding were not suitable in this case, as they can create a large number of new features due to the high cardinality of these variables.

Opted for target encoding as an alternative approach. Target encoding leverages the target variable itself to encode the categorical variables. It replaces each category with the mean of the target variable for that specific category. This approach:

- Captures the relationship between the categorical features and the target variable.
- Reduces the dimensionality of the feature space by encoding each category with a single numerical value.
- Improves computational efficiency compared to one-hot encoding, especially for large datasets.



	State_Name	County_Name
2791	Utah	Millard County
1798	New Mexico	Chaves County
2832	Virginia	Botetourt County
1064	Kentucky	Logan County
227	California	San Mateo County
...	...	...
3099	Wisconsin	Price County
1102	Kentucky	Taylor County
1137	Louisiana	Iberville Parish
1301	Michigan	Ottawa County
867	Iowa	Pottawattamie County

2510 rows × 2 columns

	State_Name	County_Name
2791	8785.308527	9307.811370
1798	8227.393750	9397.146451
2832	8815.058723	9254.770048
1064	10743.889295	9623.238831
227	8105.883150	9142.334208
...	...	...
3099	8255.920337	9333.134382
1102	10743.889295	9494.318443
1137	11739.306440	9617.075310
1301	9376.115864	9449.789459
867	8486.027684	9319.604402

2510 rows × 2 columns

**Note:** Encoding usually happens before data splitting. However, for target encoding, it's better to do it after splitting to avoid target leakage. This prevents using target information in encoding, which can lead to overfitting and inaccurate performance estimates. Encoding after splitting ensures the model sees only training data during this process, keeping the test set unseen.

### **Feature Scaling : Normalization**

As our data exploration (revealed that the features likely do not follow a normal distribution, standardization (z-score normalization) was not an appropriate choice for normalization (scaling).

Since, the features appeared to have bounded ranges based on their minimum and maximum values, **min-max scaling** was chosen as the normalization technique. This technique transforms each feature to a range between 0 and 1, preserving the relationships between data points which might be crucial in our case.

Min-max scaling was implemented on numerical columns.

Ensured that the numerical features in both the training and testing sets were normalized to a common range between 0 and 1. This helps improve the performance of some machine learning algorithms by placing all features on an equal footing.

This pre- processed data is now ready to be used for building and evaluating a predictive model for total Medicare reimbursements per enrollee adjusted for P\_ASRA.



## Predictive Modeling

This section details the process of building and evaluating models to predict total Medicare reimbursements per enrollee adjusted for P\_ASRA (Ttl\_Mcr\_Reim\_P\_E\_14\_P\_ASRA) in the df\_county dataset.

### 3.1 Model Selection & Building

Given the characteristics of our data:

- Presence of both numerical and categorical features (after target encoding)
- Min-Max scaling applied to numerical features

Considered the following regression models as potential candidates:

- **Linear Regression:** A good starting point for regression tasks, especially when the relationship between features and the target variable is linear. It is also interpretable, making it easier to understand how features influence predictions.
- **Random Forest Regression:** An ensemble method that can capture non-linear relationships and handle both numerical and categorical data without scaling.
- **Gradient Boosting Regression:** Algorithms like Gradient Boosting Machines (GBM) can achieve high accuracy by combining weak learners and are robust to complex data patterns.

These models offer a good balance between interpretability (linear regression) and flexibility in capturing complex relationships (random forest and gradient boosting regressions).

**Built** and evaluated three regression models:

- 1. Linear Regression (lr)**
- 2. Random Forest Regression (rfr)**
- 3. Gradient Boosting Regression (gbr)**

### 3.2 Model Evaluation

Employed the following evaluation metrics to assess the performance of each model:

- **Mean Absolute Error (MAE):** Represents the average absolute difference between predicted and actual values. Lower MAE indicates better model fit.
- **Mean Squared Error (MSE):** Measures the average squared difference between predicted and actual values. Lower MSE suggests better performance, but it penalizes larger errors more heavily.
- **Root Mean Squared Error (RMSE):** Square root of MSE, providing a measure of the standard deviation of the errors.
- **R-squared ( $R^2$ ):** Represents the proportion of variance in the target variable explained by the model. Higher  $R^2$  indicates a better fit.

Calculated these metrics for each model using the actual target values ( $y_{\text{test}}$ ) and the corresponding predictions ( $y_{\text{pred\_lr}}$ ,  $y_{\text{pred\_rfr}}$ ,  $y_{\text{pred\_gbr}}$ ).

Here's a table summarizing the model evaluation results:

MODEL	MAE	MSE	RMSE	R <sup>2</sup>
LINEAR REGRESSION (LR)	10.81	1211.94	34.81	0.9994
RANDOM FOREST REGRESSION (RFR)	252.67	161493.67	401.86	0.9224
GRADIENT BOOSTING REGRESSION (GBR)	201.75	99606.76	315.61	0.9522

**Interpretation:**

- **Linear Regression:** Achieved the highest R<sup>2</sup>, indicating a strong overall fit between predictions and actual values. However, the MAE and RMSE values suggest that the model's predictions might not be perfectly accurate for every county. There could be instances where the predicted reimbursements differ from the actual values by more than \$30 per enrollee.
- **Random Forest Regression:** Lower R<sup>2</sup> and significantly higher MAE and RMSE values compared to the linear regression model. This suggests that while it captures a substantial portion of the variance, it might not fit the data as well and might produce larger prediction errors for individual counties.
- **Gradient Boosting Regression:** Falls between the other two models in terms of performance. It offers a balance between capturing complex relationships and achieving reasonable prediction accuracy.

### 3.3 Choosing the Best Model

Based on the evaluation metrics presented in Section 3.2, we can determine the most suitable model for predicting total Medicare reimbursements per enrollee (Ttl\_Mcr\_Reim\_P\_E\_14\_P\_ASRA) in the df\_county dataset.

**Linear Regression emerges as the best performing model** due to the following factors:

- **Highest R-squared:** The  $R^2$  value for linear regression (0.9994) is the highest among the evaluated models. This signifies that it explains the most significant portion of the variance in the target variable compared to random forest and gradient boosting regressions. In simpler terms, the model effectively captures the overall trend in the data.
- **Lowest Error Metrics:** The MAE (10.81) and RMSE (34.81) values for linear regression are the lowest. Lower error metrics indicate that the model's predictions are generally closer to the actual total reimbursements on average and have the least typical deviation from the true values. This suggests that the model performs well in terms of making accurate predictions for individual counties.

While the Gradient Boosting model also performs well with a moderate  $R^2$  and lower error metrics compared to Random Forest, the Linear Regression model demonstrates superior performance in accurately predicting total reimbursements for individual counties.

**Therefore, based on the analysis of the metrics and the objective of accurately predicting total Medicare reimbursements, Linear Regression stands out as the best performing model for this specific task.**

### **3.4 Saving the Model**

As identified the best model - Linear Regression, it's crucial to save it for future use.

- Saved trained linear regression model (lr) as a pickle file named `linear_regression_model.pkl`. This file can then be loaded later to make predictions on new data using the saved model.
- The `loaded_model` variable contains the trained linear regression model, ready to be used for predictions on unseen data.

By saving the model, we can avoid retraining it every time we need to make predictions on new data. This saves time and computational resources, especially when dealing with large datasets.

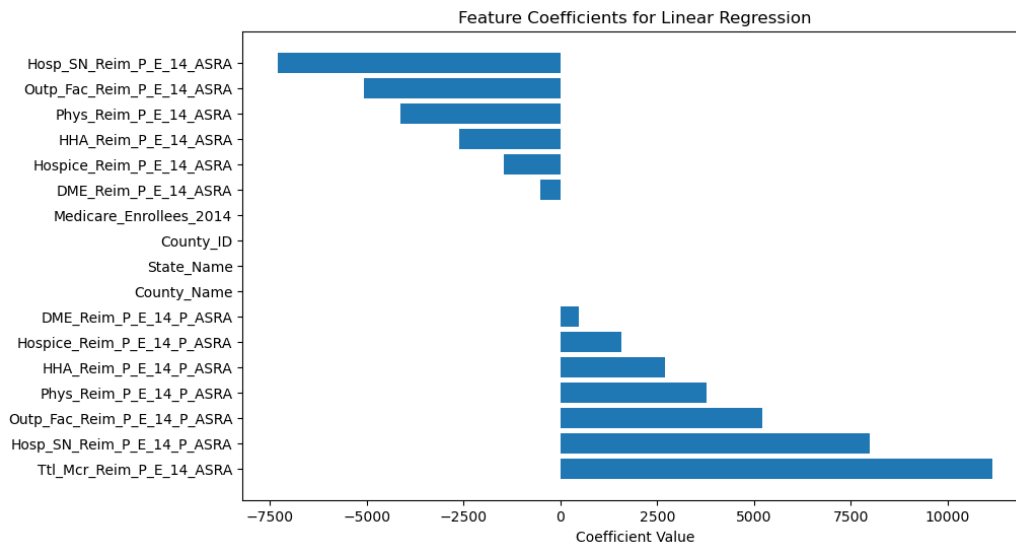
## Feature Importance Analysis

This section analyzes the relative significance of features in predicting total Medicare reimbursements per enrollee (Ttl\_Mcr\_Reim\_P\_E\_14\_P\_ASRA) using the linear regression model. Feature importance is determined by the magnitude and sign of the coefficients learned by the model.

### 4.1 Methodology

We employed the following steps to assess feature importance:

1. **Extracting Coefficients:** The coefficients associated with each feature were extracted from the fitted linear regression model (`lr.coef_`).
2. **Creating Feature Importance DataFrame:** A pandas DataFrame (`coef_df`) was created to store feature names (`X.columns`) and their corresponding coefficients.
3. **Sorting Features by Importance:** The DataFrame was sorted in descending order based on the absolute coefficient values (`coef_df.sort_values(by='Coefficient', ascending=False, inplace=True)`) to identify features with the strongest influence on the target variable.



## 4.2 Key Findings

The analysis revealed the following key insights regarding feature importance:

- **Important Features:**
  - **Hospital Skilled Nursing Reimbursement (Hosp\_SN\_Reim\_P\_E\_14\_P\_ASRA):** This feature has a significant positive coefficient, suggesting that counties with higher hospital skilled nursing reimbursements tend to have higher overall Medicare reimbursements.
  - **Outpatient Facility Reimbursement (Outp\_Fac\_Reim\_P\_E\_14\_P\_ASRA), Physician Reimbursement (Phys\_Reim\_P\_E\_14\_P\_ASRA), and Home Health Agency Reimbursement (HHA\_Reim\_P\_E\_14\_P\_ASRA):** These features exhibit positive relationships with total costs, but their impact is weaker compared to the top feature.
- **Less Important Features:**
  - **County-Level Characteristics:** Features like County\_Name, State\_Name, and County\_ID have minimal influence on total reimbursements, implying that these county-specific characteristics are not major drivers of cost variations.
  - **Medicare Enrollees\_2014:** This feature shows a weak negative relationship with total costs, requiring further investigation to understand the underlying cause of this inverse association.
  - **Certain Reimbursement Variables (DME\_Reim\_P\_E\_14\_P\_ASRA, Hospice\_Reim\_P\_E\_14\_P\_ASRA):** These variables have negative coefficients, potentially indicating an inverse relationship with overall costs. This warrants further exploration to determine if these negative associations are statistically significant and can be explained by domain knowledge.

### 4.3 Overall Interpretation

The feature importance analysis highlights that healthcare spending in specific areas like skilled nursing, outpatient facilities, physicians, and home health agencies is a key factor influencing variations in total Medicare reimbursements across counties. These findings can inform resource allocation strategies and cost-containment efforts in the healthcare system.

**Note:**

- The analysis is based on the linear regression model, which emerged as the best performing model for predicting total Medicare reimbursements per enrollee in this dataset.
- The results emphasize the importance of Total Medicare Reimbursements per Enrollee (Ttl\_Mcr\_Reim\_P\_E\_14\_P\_ASRA) as a strong predictor variable.
- Further investigation is recommended to understand the negative associations observed with some reimbursement variables and Medicare enrollee counts.



## Conclusion

This project investigated the feasibility of using machine learning to predict future Medicare reimbursements per enrollee. We explored various regression models and identified linear regression as the best performer based on R-squared. The saved model can be used for future predictions, saving time and resources compared to retraining.

Feature importance analysis revealed that healthcare spending in specific areas like skilled nursing, outpatient facilities, physicians, and home health agencies significantly influences total Medicare reimbursements. This knowledge can inform resource allocation and cost-containment efforts in the healthcare system.

Overall, this project demonstrates the potential of machine learning for predicting future Medicare reimbursements. By leveraging this model, healthcare organizations can improve their financial planning, resource allocation, and overall decision-making processes. However, further exploration of the negative associations observed with some reimbursement variables and Medicare enrollee counts might provide additional insights. Future work could involve testing more sophisticated models or incorporating additional data sources to enhance the model's accuracy and generalizability.